KA³: Weiterentwicklung von Sprachtechnologien im Kontext der Oral History

Köhler, Joachim; Gref, Michael; Leh, Almut

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
Verlag Barbara Budrich

Empfohlene Zitierung / Suggested Citation:
Köhler, J., Gref, M., & Leh, A. (2017). KA³: Weiterentwicklung von Sprachtechnologien im Kontext der Oral History. *BIOS - Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen*, *30*(1-2), 44-59. https://doi.org/10.3224/bios.v30i1-2.05

KA³

Weiterentwicklung von Sprachtechnologien im Kontext der Oral History

Joachim Köhler, Michael Gref und Almut Leh

1. Oral History-Interviews als Quellen biographischer Forschung

Die Befragung von Zeitzeugen und damit das Interesse an biographischen Verläufen und subjektiven Selbstauskünften hat in den Sozial- und Geisteswissenschaften eine lange Tradition. In der Soziologie gilt die Studie *The Polish Peasant in Europe and America* des Chicagoer Soziologen Isaac Thomas und seines polnischen Kollegen Florian Znanieck, erschienen 1918/1920, als Ausgangspunkt für die Entwicklung biographischer Methoden. Und in der Geschichtswissenschaft werden mündliche Erinnerungszeugnisse gar seit der Antike als Quellen genutzt.

Entscheidenden Aufschwung nahm die Befragung von Zeitzeugen in den 1970er und 1980er Jahren, als in nahezu allen Humanwissenschaften ein zunehmendes Interesse an biographischer Forschung entstand, namentlich in der Soziologie und Pädagogik, in der Volkskunde und Ethnologie, in der Geschichtswissenschaft und Literaturwissenschaft, in Psychoanalyse und Psychologie. Wesentliche Voraussetzungen für die Durchführung von Zeitzeugenbefragungen waren technischer Natur: die Entwicklung einer handlichen und bald auch preisgünstig verfügbaren Aufnahmetechnik in Gestalt des Kassettenrecorders. Diese einfach zu handhabende Technik ermöglichte es sogar Laien, Oral History-Interviews durchzuführen. Unter der Vorstellung, Interviews mit Zeitzeugen böten einen unmittelbarer Zugang zur Geschichte, erfreute sich die Oral History tatsächlich vor allem in Geschichtswerkstätten sowie in schulischen und außerschulischen Bildungsprojekten großer Beliebtheit, bevor sie sich mit der interviewbasierten Studie „Lebensgeschichte und Sozialkultur im Ruhrgebiet 1930 bis 1960“ Anfang der 1980er Jahre auch im universitären Bereich Anerkennung verschaffen konnte Niethammer 1983a; 1983 b; Niethammer/von Plato 1985).

Seither wurde unter zeithistorischen Fragestellungen eine Vielzahl von universitären und außeruniversitären Projekten durchgeführt, in denen Zeitzeugeninterviews erhoben und ausgewertet wurden. Thematische Schwerpunkte waren vor allem die Zeit des Nationalsozialismus und des Zweiten Weltkrieges. Doch auch zu vielen anderen Themen und historischen Phasen sind Interviews geführt worden, so dass in den vergangenen vier Jahrzehnten mehrere tausend Zeitzeugen zu unterschiedlichen Aspekten der Geschichte des 20. Jahrhunderts befragt wurden. Die materielle Hinterlassenschaft dieser Projekte sind tausende von Audiokassetten, inzwischen auch eine unüberschaubare Menge von Videoaufzeichnungen auf unterschiedlichen Datenträgern, an denen sich die Technikgeschichte vergangener Jahrzehnte nachverfolgen lässt.

Glücklicherweise war zumindest im universitären Kontext teilweise die Vorstellung handlungsleitend, dass es sich bei Zeitzeugeninterviews um Quellen handelt, die nicht

 https://doi.org/10.3224/bios.v30i1-2.05

nur zum Zweck der Überprüfbarkeit der Forschungsergebnisse archiviert werden sollten, sondern vor allem auch, um für zukünftige Forschungen unter eventuell neuen Fragestellungen mit Gewinn analysiert werden zu können. Unter dieser Perspektive wurde 1994 an der FernUniversität in Hagen das Archiv „Deutsches Gedächtnis" gegründet, in dem heute rund 3.000 Zeitzeugeninterviews archiviert sind, die aus über einhundert überwiegend historisch ausgerichteten Forschungsprojekten stammen, darunter auch die des Pionierprojektes „Lebensgeschichte und Sozialkultur im Ruhrgebiet".[1] Ein weiteres umfangreiches Oral History-Archiv ist die etwa zeitgleich entstandene Werkstatt der Erinnerung an der Forschungsstelle für Zeitgeschichte in Hamburg, wo Interviews mit Bezug zum norddeutschen Raum archiviert werden und für Forschungen zur Verfügung stehen (Apel 2011).

Dass sich diese Art Interviews für weitere Auswertungen anbieten, hängt nicht zuletzt mit dem Interviewformat zusammen, das sich in der Oral History herausgebildet hat und ein vielschichtiges qualitatives Material hervorbringt, dessen Potential mit einer einzigen Untersuchung kaum ausgeschöpft werden kann. In Anlehnung an die soziologische Biographieforschung werden Zeitzeugen zumeist nach der von Fritz Schütze entwickelten Methode des narrativen Interviews befragt (Schütze 1976). Charakteristisch für diese Methode ist, dass das Interview nicht durch Fragen strukturiert wird, sondern der Interviewpartner aufgefordert wird, ausführlich und nach eigenen Relevanzkriterien seine Lebensgeschichte zu erzählen. In allen Phasen ist die Interviewführung darauf ausgerichtet, unvorbereitete Stegreiferzählungen von Geschehensverläufen hervorzulocken, an denen der Erzähler aktiv oder passiv beteiligt war. So entstehen mehrstündige Interviewaufzeichnungen, die die gesamte Biographie des Interviewten umspannen und in freier Erzählung eine Vielzahl von Themen berühren. Meist werden mit der Erstauswertung im Nachgang der Erhebung nur bestimmte Aspekte und Themen analysiert, während vieles ausgespart bleibt, was im Rahmen anderer Fragestellungen mit Gewinn erforscht werden könnte.

Tatsächlich erleben wir heute ein stetig zunehmendes Interesse an der Sekundärauswertung von Interviews, die in früheren Zeiten und teils unter anderen Fragestellungen geführt wurden und in Archiven zugänglich sind.[2] Dieser Trend zur Sekundäranalyse dürfte mehrere Ursachen haben. Zum einen kommt in der Nutzung von Zeitzeugeninterviews die inzwischen erreichte Akzeptanz dieser Quelle innerhalb der historischen Zunft zum Ausdruck. Heute ist es für eine seriöse historische Studie bei entsprechenden Themen unabdingbar, neben schriftlichen Dokumenten auch die Perspektive von Zeitzeugen zu berücksichtigen – sofern entsprechende Quellen wie Audio- oder Videointerviews in Archiven vorliegen. Und aus historischen Ausstellungen, Museen und Gedenkstätten sind Zeitzeugeninterviews gar nicht mehr wegzudenken.

Eine zweite Ursache für vermehrte Reanalysen ist die große Anzahl von Interviews, die in den vergangenen Jahrzehnten geführt wurde. Es ist oftmals gar nicht erforderlich, selbst Interviews zu führen; in vielen Fällen ist dies allerdings auch gar nicht mehr möglich. Die Erforschung des Nationalsozialismus und des Zweiten Weltkrieges ist dafür ein gutes Beispiel. Denn während das Interesse der Geschichtswissenschaft ebenso wie der politischen Öffentlichkeit an diesen Themen unvermindert anhält, stehen Menschen, die diese Phase der Geschichte bewusst erlebt haben, kaum noch für

1 Weiteres zu den Sammlungen im Archiv „Deutsches Gedächtnis" siehe Leh 2015.

2 Vgl. zur Sekundärauswertung Apel 2015.

Befragungen zur Verfügung. Archivierte Interviews aus früheren Projekten werden deshalb schon bald der einzige Zugang zu den Erfahrungen und Erinnerungen dieser Generation sein.

Das wachsende Interesse an Zeitzeugeninterviews manifestiert sich auch in einem anderen Trend. Parallel zur Nutzung etablierter Oral History-Archive werden aktuell eine Vielzahl von Interviewprojekten durchgeführt, deren Ziel allein die Dokumentation der Lebensgeschichten ist. Während in früheren Oral History-Projekten Interviewerhebung und Auswertung unter historischen Fragestellungen eng zusammengehörten, konzentrieren sich diese Projekte auf die Produktion der Quelle, wobei der Aspekt der Bewahrung vor dem Vergessen angesichts des Sterbens der Zeitzeugen stark gemacht wird. Eine etwaige Auswertung wird späteren Forschungen anheimgestellt. Typisch für diese Dokumentationsprojekte ist die Präsentation der Interviews in einem Internetportal. Stilbildend war hier möglicherweise das von Steven Spielberg initiierte *Visual History Archive* des Shoah Foundation Institute, das mit 52.000 Interviews mit Überlebenden und Zeugen des Holocaust weltweit größte Archiv mit videographierten Oral History-Interviews (Leh/Tausendfreund 2011, Pagenstecher in diesem Heft). In der Folge entstanden unter anderem das Mauthausen Survivors Documentation Project mit 850 Interviews[3] und das Archiv „Zwangsarbeit 1939-1945. Erinnerungen und Geschichte“ mit rund 600 Interviews (Näheres s. Pagenstecher in diesem Heft).[4] Von den aktuellen Interviewprojekten sei hier stellvertretend nur die von der Stiftung Geschichte des Ruhrgebiets in Kooperation mit dem Deutschen Bergbau-Museum Bochum durchgeführte Befragung von rund einhundert „Menschen im Bergbau“ genannt, deren Interviews in einem Online-Portal zugänglich sind.[5]

Bemerkenswert und folgenreich ist, dass diese Dokumentationsprojekte nicht nur auf die Forschung verzichten, sondern vielfach auch auf die Transkription der Interviews. So positiv die Verfügbarmachung der Originalquelle in Form der Videoaufzeichnung ist, so problematisch ist der Verzicht auf die Transkription. Zweifellos ist die Verschriftlichung sehr zeitaufwändig und damit ein erheblicher Kostenfaktor in der Projektkalkulation. Die begrenzten Mittel in die Durchführung weiterer Interviews zu investieren, statt Teile für Transkriptionen vorzusehen, ist eine nachvollziehbare Entscheidung, wird aber zumindest mittelfristig die Nutzbarkeit der Interviewquellen einschränken. Sowohl für die Analyse von Oral History-Interviews wie auch für deren Archivierung ist das Transkript ein unverzichtbares Hilfsmittel. In der Analyse ist es ein Instrument der kritischen Distanznahme zum Quellenmaterial und der besseren Handhabung der Informationsfülle. Bei Veröffentlichungen in konventionellen Printmedien ist das Transkript unerlässlich für das Zitieren von Belegstellen. In der Archivpraxis ist die Textfassung des Interviews derzeit das wichtigste Instrument bei der Recherche nach relevanten Interviews oder Interviewpassagen für Sekundäranalysen. Tatsächlich ist das Wiederauffinden bestimmter Inhalte angesichts der qualitativen Daten und deren überbordender Menge die größte Herausforderung an Oral History-Archive. Ohne effektive Suchstrategien droht einer Vielzahl von Interviews das Vergessen. Eine ebenfalls zeitaufwändige Verschlagwortung kann die Vielfalt der Forschungsfragen

3 https://msrp.univie.ac.at/project-information/msdp/.

4 http://www.zwangsarbeit-archiv.de/.

5 https://menschen-im-bergbau.de/. Weitere aktuelle Dokumentationsprojekte bei Apel (2015: 245).

nur unzureichend abbilden. Die derzeit erfolgreichste Strategie ist immer noch die Volltextsuche nach einschlägigen Begriffen über die Gesamtheit der Transkripte. Das Problem sind deshalb die nicht-transkribierten Interviews in den Oral History-Archiven und in den derzeit entstehenden Dokumentationsprojekten und Online-Portalen. Das Interesse an einer automatisierten Spracherkennung, mit der Transkripte und Schlagwörter generiert und Begriffe direkt im Audiosignal gesucht werden können, ist in der interviewbasierten Forschung und bei einschlägigen Archiven einfach riesig.

Schaut man sich in den Digital Humanties um, stellt man fest, dass die hier diskutierten Entwicklungen sich vor allem auf Text- und Bilddokumente beziehen; der Bereich audiovisueller Daten spielt bisher eine nachgeordnete Rolle. Und während die alltägliche Gegenwart von Siri, Alexa und Co den Eindruck vermittelt, eine leistungsfähige Spracherkennung sei längst Stand der Technik, ist die Erfahrung in der Forschungspraxis eine deutlich andere. Tatsächlich stellen biographische Interviews Sprachtechnologien noch immer vor eine große Herausforderung. Im Projekt KA³ wird dieser Anwendungsfall erstmals systematisch analysiert und an Verbesserungen gearbeitet.[6]

2. Technologien zur Sprachanalyse und Spracherkennung

In den letzten Jahren wurden auf dem Gebiet der Spracherkennung enorme technologische Fortschritte erzielt. Mittlerweile gibt es eine Reihe von leistungsfähigen Spracherkennungssystemen in Form von Sprachassistenten und Spracheingabefunktionen bei Smartphone-Apps. Durch die weite Verbreitung von intelligenten Lautsprecherboxen, wie Amazon Echo oder Google Now haben Sprachassistenten den Einzug in die Wohnungen der Nutzer gefunden. Die intelligenten Lautsprecherboxen werden über ein Schlüsselwort aktiviert, um die Aufzeichnung und Erkennung der Sprache des Benutzers zu starten. Anschließend werden die transformierten und verschlüsselten Sprachdaten über das Internet auf leistungsfähige Serversysteme übertragen, die dann die eigentliche Erkennung vornehmen. Die Spracherkennung erfolgt somit in der Cloud. Die Ausgabeergebnisse werden entweder als reiner Text oder als vertonte Antwort zurückgesendet und über die Lautsprecherbox ausgegeben.

Die eigentliche Spracherkennung erfolgt mittels statistischer Verfahren. Das aufgenommene Sprachsignal besteht aus einer Folge von digitalen Abtastwerten, die in spektrale Merkmale, beispielsweise sogenannte „Filter-Bank-Features“ oder „Mel-Frequency-Cepstrum-Coefficients“ (kurz MFCC) umgewandelt werden. Diese Merkmale werden verwendet, um die Wahrscheinlichkeit von unterschiedlichen Sprachlauten zu

6 Das Projekt „Kölner Zentrum Analyse und Archivierung von AV-Daten. KA3“ dient dem Aufbau und der Weiterentwicklung eines fach- und standortübergreifenden Kölner Zentrums für Analyse und Archivierung audiovisueller Daten (KA³) mit den drei Komponenten Analyse, Archivierung/Publikation und Schulung/Beratung. Besondere Aufmerksamkeit gilt den miteinander zusammenhängenden Problemen der interaktionsbezogenen Strukturierung und der effizienten Bereitstellung und Archivierung von audiovisuellen Daten, die sowohl geisteswissenschaftlich wie informationstechnologisch erforscht werden sollen und die von grundlegender Bedeutung für die wissenschaftliche Arbeit mit AV-Daten sind. Das Projekt wird gefördert vom Bundesministerium für Bildung und Forschung und koordiniert von Nikolaus P. Himmelmann, Institut für Linguistik, Universität zu Köln. Joachim Köhler und Michael Gref (Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS) sind im Projekt für die Weiterentwicklung von Sprachtechnologen verantwortlich, Almut Leh (FernUniversität in Hagen) für Bedarfsanalysen, Datenbereitstellung und Evaluation.

jedem Zeitpunkt des Sprachsignals zu schätzen. Diese Wahrscheinlichkeiten der einzelnen Sprachlaute wurden lange mit sogenannten Hidden-Markov-Modellen und Gauß'schen Mischverteilungen ermittelt. Seit einigen Jahren werden zur statistischen Modellierung der Sprachlaute künstliche neuronale Netze mit verschiedensten Architekturen verwendet. Diese beiden Verfahren sind Methoden aus dem Gebiet des Maschinellen Lernens und müssen mit Daten auf die jeweilige Aufgabenstellung „trainiert" werden.

Zunächst wurden mehrschichtige, vollverbundene Neuronale Netze verwendet (sogenannte Deep Neural Networks), die bereits zu einer deutlich besseren Erkennungsleistung führen als gewöhnliche Gauß-Mischverteilungen. Mittlerweile werden fortschrittlichere Arten von neuronalen Netzen eingesetzt, wie sogenannte Long-Short-Term-Memory-Networks (Hochreiter und Schmidhuber 1997), (Sak, Senior und Beaufays 2014), Time-Delay Neural Networks (Waibel, et al. 1989), (Peddinti, Povey und Khudanpur 2015) und Netze, die verschiedene grundlegende Netzarten in einem Netz kombinieren (Cheng, et al. 2017). Diese Netze sind dank ihres rekurrenten Aufbaus in der Lage, den zeitlichen Verlauf der gesprochenen Sprache exakter zu modellieren. Die Gesamtheit der Verfahren zur Modellierung und Erkennung der Lautwahrscheinlichkeiten wird als akustische Modellierung bezeichnet. Für die Erstellung der statistischen Lautmodelle, auch als „akustische Modelle" bezeichnet, werden sehr umfangreiche annotierte Sprachdatenbanken benötigt. Diese Daten enthalten eine exakte Worttranskription auf Segmentebene. In der Regel gilt: Je mehr annotierte Sprachdaten vorliegen, desto leistungsfähiger werden die Modelle, die auf diesen Daten trainiert werden. Das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) verwendet für die Spracherkennung das open-source-System Kaldi (Povey, et al. 2011), das in starkem Maße die neuronalen Netze einsetzt. In dem aktuellen Fraunhofer IAIS-Spracherkennungssystem wird das akustische Modell mit 1.000 Stunden annotierter Sprachdaten aus dem Rundfunkbereich trainiert (Stadtschnitzer, et al. 2014).

Neben der akustischen Modellierung gilt es, die Grammatik bzw. die Folge der gesprochenen Wörter zu modellieren und zu berechnen. Dazu werden sogenannte Sprachmodelle erstellt, die die Wortfolgewahrscheinlichkeiten modellieren. Diese Modelle enthalten die Wahrscheinlichkeiten von unterschiedlichen Wortfolgen. Diese Wortfolgen haben in gewöhnlichen Sprachmodellen in der Regel eine Länge von drei bis fünf Wörtern. Um diese Vielzahl von Modellparametern zuverlässig zu bestimmen, werden ebenfalls sehr große Textmengen verwendet. Oftmals werden mehr als 100 Millionen fortlaufende Wörter herangezogen, um die Wortfolgewahrscheinlichkeiten zu bestimmen. Textdaten aus unterschiedlichen Domänen können sich hier in hohem Maße auf das resultierende Erkennungsergebnis auswirken. So weisen beispielsweise wissenschaftliche Veröffentlichungen andere Formulierungen und Wortfolgenhäufigkeiten auf als Artikel aus dem Sportbereich.

Die Verbindung zwischen Sprachlauten und Wortfolgewahrscheinlichkeiten erfolgt über das phonetische Aussprachelexikon. Dieses enthält für jedes Wort eine phonetische Transkription, die entweder manuell über eine genaue phonetische Transkription oder über ein automatisches Verfahren generiert wird. Für das Wort „Kriegskinder" lautet beispielsweise die phonetische Transkription „k r i: k s k ɪ n d ɐ".

Mit den ermittelten akustischen Phonemwahrscheinlichkeiten, den Informationen zur Aussprache einzelner Wörter sowie den Wortfolgewahrscheinlichkeiten wird die

automatische Spracherkennung durchgeführt. Somit handelt es sich bei der Spracherkennung um einen statistischen Prozess, der die wahrscheinlichste Wortfolge, basierend auf Beobachtung der akustischen Merkmale und den vorberechneten Modellen, ausgibt. Natürlich besteht die Erwartung, dass der ausgegebene Text möglichst fehlerfrei ist. Allerdings können die Erkennungsraten von typischen Erkennungssystemen bei unterschiedlichen Aufgabenstellungen sehr stark variieren.

3. Das Fraunhofer IAIS Audio Mining-System

Das Fraunhofer IAIS Audio Mining-System (Schmidt, Stadtschnitzer und Köhler 2016) wurde ursprünglich für die automatisierte Erschließung zahlreicher und sehr umfangreicher Mediadaten von audiovisuellen Beiträgen aus dem Rundfunkbereich entwickelt. Diese zeichnen sich in der Regel durch eine hohe Sprachqualität hinsichtlich Sprechweise (Aussprache) und Tonqualität (Aufnahmetechnik) aus. Da die Rundfunksender über äußerst umfangreiche Archive und AV-Bestände verfügen, die über manuelle Prozesse nicht mehr erschließbar sind, besteht erhöhter Bedarf, die Sendungen mittels automatischer Technologien und Systeme zu erschließen und eine Recherche in der Tonspur zu ermöglichen. Diese Technologie zur Analyse von Sprachdaten wird in Anlehnung an das Text Mining für textuelle Daten als Audio Mining bezeichnet. Audio Mining-Technologien ermöglichen den inhaltlichen Zugriff auf die gesprochenen Daten.

Das Fraunhofer Audio Mining System besteht aus mehreren Komponenten, deren Zusammenspiel eine möglichst genaue Erschließung und Transkription der Sprachdaten herbeiführen soll. Die Verarbeitung von Sprachsignalen durch diese Komponenten lässt sich entsprechend in verschiedene, aufeinanderfolgende Schritte gliedern. Im ersten Schritt, der sogenannten Diarisierung, wird die Sprachdatei in homogene Abschnitte zerlegt mit dem Ziel, Abschnitte einzelner Sprecher zu erhalten. Mittels eines unüberwachten Algorithmus (Tritschler und Gopinath 1999) werden Sprecherwechsel automatisch erkannt, falls die Abschnitte eines einzelnen Sprechers mindestens fünf Sekunden lang sind. Diese Sprachabschnitte werden dann, jeweils einzeln, dem eigentlichen Spracherkennungsmodul zugeführt, das im Folgenden die Umwandlung des Sprachsignal in Text vornimmt. Über die letzten Jahre wurden deutliche Verbesserungen bei der Spracherkennungsqualität erzielt. So konnte die Wortfehlerrate für den Anwendungsbereich Rundfunk (Baum, et al. 2010) von über 25% auf 8% reduziert werden. Dies kann als großer Fortschritt bei der Verbesserung von Sprachtechnologien bezeichnet werden.

Nach der durchgeführten Spracherkennung werden die vollständigen Transkripte mittels eines Schlüsselwortextraktors weiterverarbeitet. Basierend auf dem tf-idf (Term Frequency – Inverse Document Frequency)-Verfahren, werden die wichtigsten Schlüsselwörter des Transkripts für einen Beitrag ausgegeben. Die Schlüsselwörter geben eine erste Orientierung über Themen und Inhalte eines Beitrages. Die Schlüsselworterkennung mittels tf-idf berücksichtigt die Häufigkeit der Wörter in der deutschen Sprache und die relative Häufigkeit in dem automatisch generierten Transkript. So werden seltener vorkommende Wörter der deutschen Sprache (z.B. Eigennamen) eher als Schlüsselwort ausgewählt als häufig genannte Wörter (z.B. „Sie“, „Prozent“, „Millionen“).

Sämtliche automatisch erzeugten Analyseergebnisse werden mit den entsprechenden Zeitinformationen in einer umfassenden Metadatendatei zusammengefügt und abgespeichert. Als Datenformat wurde das MPEG-7-Format gewählt, das ein standardisiertes XML-Schema für audiovisuelle Metadaten darstellt.

Die Module zur Sprechersegmentierung, Spracherkennung und Schlüsselwortgenerierung sind in der C++ Softwarekomponente iFinder zusammengefasst. Dieses Analysesystem ist wiederum in eine Workflowumgebung integriert, die es erlaubt, skalierbar umfangreiche audiovisuelle Datenmengen parallel zu verarbeiten. Die MPEG-7-Metadaten werden nach der Verarbeitung persistiert, d.h. mit einem eindeutigen und dauerhaften Identifikator ausgezeichnet und abgespeichert. Die erzeugten Metadaten (u.a. Transkripte) werden mittels der Suchmaschine Solr indiziert, so dass komfortabel nach Inhalten gesucht werden kann. Die Verarbeitungsaufträge können über eine webbasierte Schnittstelle an das Analysesystem übergeben werden.

Die Gesamtheit der oben genannten Komponenten stellt die Fraunhofer IAIS Audio Mining-Lösung dar. Diese wird durch eine webbasierte grafische Benutzerschnittstelle ergänzt. Nachfolgender Screenshot (Abbildung 1) stellt die wesentlichen Ausgabe- und Navigationselemente dar.

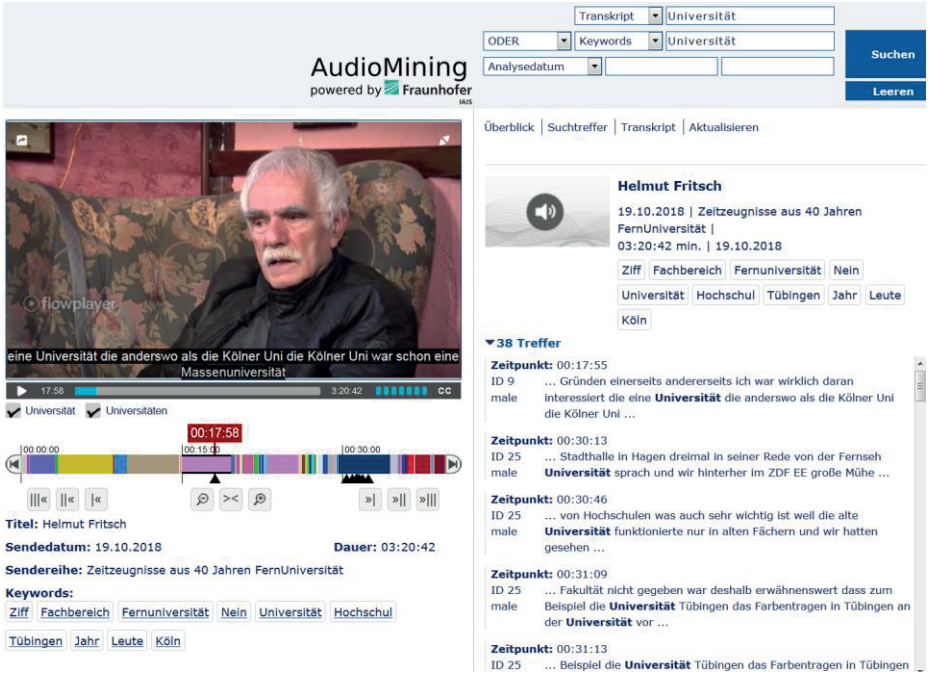

Abbildung 1: Such- und Rechercheoberfläche des Fraunhofer IAIS Audio Mining Systems mit einem Oral History-Interview

Auf der linken Seite der Benutzeroberfläche befindet sich ein Media-Player, der analysierte Mediadateien wiedergibt und dabei Untertitel auf Basis des automatisch erzeugten Transkripts anzeigt. Unterschiedliche Sprecher werden durch verschiedene Farben in der Zeitleiste unter dem Media-Player repräsentiert. Diese Leiste erlaubt Nutzerinnen

und Nutzern schnell zwischen Segmenten von Sprechern zu wechseln und alle Segmente eines Sprechers/einer Sprecherin direkt anzusteuern. Die Suchmaske in der oberen rechten Ecke erlaubt die Eingabe von Suchbegriffen. Da alle indexierten Wörter aus den automatisch generierten Transkriptionen mit Zeitmarken versehen sind, kann jedes Suchwort in der Tonspur angesteuert und angezeigt werden. Die Sucheingaben können durch „und“ bzw. „oder“-Verknüpfungen modifiziert werden. Darüber hinaus kann ein zeitlicher Bereich angegeben werden, in dem zwei Suchwörter aufeinander folgen, so dass die gesuchten Inhalte eingegrenzt werden und die Treffergenauigkeit zunimmt. Darüber hinaus erlaubt die Oberfläche des Fraunhofer IAIS Audio Mining Systems den Export der vorliegenden MPEG-7-Metadaten. Diese lassen sich dann in andere Formate (z.B. ELAN oder txt-Dateien) abspeichern und gegebenenfalls mit anderen Werkzeugen bearbeiten.

Der Anwender kann über eine Upload-Funktionalität eigene Interviewdaten hochladen und durch das System verarbeiten lassen. Dazu kann der Anwender eine Audiodatei in den gängigen Audioformaten (z.B. mp3, wav) über sein Dateisystem markieren und in die Audio Mining Oberfläche ziehen. Die Verarbeitung dauert in der Regel so lange wie die Länge der Audiodatei. So beträgt die Verarbeitung eines dreistündigen Interviews ebenfalls drei Stunden. Anschließend ist das Interview in dem Audio Mining-System gespeichert und kann über die automatisch generierten Metadaten recherchiert und angezeigt werden.

4. Weiterentwicklungen für die Oral History im Projekt KA³

Im Rahmen des durch das Bundesministerium für Bildung und Forschung geförderten Projektes „Kölner Zentrum Analyse und Archivierung von AV-Daten - KA³“ wird das Audio Mining-System für den Anwendungsbereich der interviewbasierten Forschung angepasst und weiterentwickelt. Ziel des Projektes ist die Bereitstellung von Technologien und Diensten für die Geistes- und Kulturwissenschaften. Das Kölner Zentrum hat zum Ziel, vor allem Dienste für die Erschließung und Analyse von audiovisuellen Daten bereitzustellen. Als Anwendungsfälle wurden zwei Szenarien ausgewählt. Das Interaktionsszenario repräsentiert die Anwendung im Bereich der Sprachwissenschaften. Hier werden vor allem Interaktionsverläufe analysiert. Datenbasis sind Sprachaufnahmen, in denen sich zwei Sprecher natürlich unterhalten und es häufig zu sogenannten Backchannel-Ereignissen kommt. Aus der Analyse sollen Interaktionsmuster abgeleitet werden.

Im Interviewszenario werden Oral History-Interviews mit dem Fraunhofer Audio Mining-System verarbeitet. Dabei werden mehrere Ziele verfolgt. Zunächst soll der Transkriptionsaufwand deutlich gesenkt werden. Aktuell beträgt der Aufwand für die manuelle Transkription ungefähr die 10 bis 15-fache Zeit der Länge des Interviews. Bei einem Interview von zwei Stunden werden daher 20 und mehr Stunden für die Verschriftlichung benötigt. Ein weiterer Vorteil bei der Verwendung der Fraunhofer IAIS Audio Mining-Lösung liegt in den interaktiven Abruffunktionalitäten. Bestimmte Suchbegriffe können über die Benutzerschnittstelle schnell angesprungen werden, und das Interview lässt sich sehr viel einfacher und genauer durchsuchen. Durch das Vorhandensein von Zeitmarken der gesprochenen Wörter verschwindet der Medienbruch zwischen dem gedruckten Text und der Tonaufzeichnung. Ein weiteres Ziel bei der

Verwendung der Sprachanalysetechnologien besteht in dem Auffinden von Interaktionsmustern, ähnlich dem oben genannten Interaktionsszenario. Hierzu sollen Sprecherwechsel, Nachfragen des Interviewers, Antworten des Interviewten genauer analysiert werden.

Für das Interviewszenario stellt das Archiv „Deutsches Gedächtnis" an der FernUniversität in Hagen seine Interviewbestände zur Verfügung, begleitet den Entwicklungsprozess mit Evaluationen und testet in Pilotprojekten die Anwendung der Sprachtechnologien für die interviewbasierte biographische Forschung. Die Interviewbestände des Archivs „Deutsches Gedächtnis" sind besonders geeignet, weil sie aus einer Vielzahl von Projekte aus verschiedenen Disziplinen und Zeitläufen stammen und somit nicht nur eine Vielzahl von Sprechweisen, Dialekten und Slangs, sondern auch unterschiedliche Interviewsettings und Aufnahmetechniken repräsentieren und dabei vielfältige Themen adressieren.

Die oftmals ursprünglich analog aufgezeichneten Audiointerviews wurden inzwischen alle digitalisiert. Die Videointerviews, die auf unterschiedlichen Medien aufgezeichnet wurden, sind bisher nur teilweise digitalisiert. Alle Interviews sind mit Metadaten versehen und in einer Datenbank verzeichnet, so dass die Bestände nach formalen Kriterien wie Geburtsjahr oder -ort, Projekt oder Archivierungsstatus durchsuchbar sind. Inhaltliche Recherchen erfordern jedoch Transkripte, die nur für gut die Hälfte der Interviews vorhanden sind. Nur zehn Prozent der Interviews sind überdies zeitcodiert, d.h. die Transkripte sind mit Zeitstempeln versehen, so dass Transkript und Ton bzw. Video synchron, beispielsweise als Untertitel, wiedergegeben werden können.

Um die Leistungsfähigkeit des Audio Mining-Systems für Oral History bewerten zu können, wird eine repräsentative Testmenge an Aufzeichnungen aus dem Archiv ausgewählt. Die Vielfalt der Interviewbestände im „Deutschen Gedächtnis" erlaubt es, in den Testverfahren unterschiedliche Aufnahmetechniken, Interviewformate, Dialekte und Aussprachen zu berücksichtigen. Interviews, die durch Aufnahmequalität oder Aussprache auch für das menschliche Ohr unverständlich sind, wurden nicht in die Tests einbezogen. Im Übrigen repräsentiert die Testauswahl frühe und neuere Aufzeichnungen, so dass Alterungsprozesse der Aufnahmemedien berücksichtigt werden. Nach Alter und Geschlecht entsprechen die Testdaten dem Gesamtbestand. Im Hinblick auf Interviewmethoden und -settings wurden Interviews aus unterschiedlichen Disziplinen ausgewählt. Insgesamt umfasst die ausgewählte Testmenge 3,5 Stunden Sprachaufzeichnungen von 35 verschiedenen Sprechern (vgl. Gref/Köhler/Leh 2018).

5. Welchen Nutzen hat das Audio Mining für die interviewbasierte Forschung?

Das Audio Mining ist sowohl für Archive, die Zeitzeugeninterviews archivieren, als auch für Forschende, die Interviews als Daten bzw. Quellen verwenden, von großem Wert. In der Archivpraxis ermöglichen diese Werkzeuge die Recherche in großen Datenmengen bei direktem Zugriff auf das Audiosignal und Vorstrukturierung des Inhaltes, so dass aus der Vielzahl von Interviews gezielt Daten zu Rechercheanfragen bereitgestellt werden können. Durch die Spracherkennung können nicht-transkribierter Interviews (auch in großen Interviewarchiven bis zu 50 Prozent der Bestände) einbezogen werden, die somit für Sekundäranalysen genutzt werden können, während sie ohne Spracherkennung für weitere Forschungen verloren wären. Tatsächlich ist es finanziell

nicht darstellbar, Interviews aus abgeschlossenen Projekten nachträglich zu transkribieren.

Auch die automatisch generierten Schlüsselwörter sind in der Archivpraxis ein Fortschritt bei der Auffindung relevanter Interviews bzw. Interviewpassagen unabhängig vom Transkriptionsstatus. Dabei steigert die im Audio Mining angebotene Trefferauswertung, mit deren Hilfe die Relevanz eines Treffers unmittelbar beurteilt werden kann, die Effizienz der Recherche. Angesichts der aktuellen Fehlerraten bei der Spracherkennung sind die Treffer bei Begriffssuche und Schlüsselwörtern unvollständig, aber schon jetzt ein deutlicher Mehrwert gegenüber der Beschränkung auf transkribierte Interviews.

Auch in der Analyse ist der direkte Zugriff auf das Audiosignal für Oral History und Biographieforschung ein erheblicher Gewinn, geradezu die Umsetzung des bisher selten eingelösten Anspruchs, die Audioaufzeichnung als Primärquelle zu behandeln und damit die Art des Sprechens (Sprechmelodie, Stimmqualität, Pausen etc.) in die Interpretation einzubeziehen. Tatsächlich beruht die Interviewanalyse vielfach allein auf der Kenntnis des Transkriptes, was vor allem bei der Auswertung nicht selbst geführter Interviews Ursache für Fehlinterpretationen sein kann. Demgegenüber ermöglicht der Zugriff auf das Audiosignal die synchrone Darstellung von Audio und Transkription und damit die Rezeption der Primärquelle als Voraussetzung einer der Aussageabsicht angemessenen Interpretation des Interviews.

Darüber hinaus eröffnen die Werkzeuge neue Dimensionen für Forschungsfragen. Während die Fallzahlen bei konventionellen Analysemethoden meist bei um die 30 Interviews liegen, können mit technischer Unterstützung viel größere Fallzahlen bearbeitet und somit unter vielfältigen vergleichenden Fragestellungen auch quantitativ ausgewertet werden. Gleichzeitig bieten die Werkzeuge auch der qualitativen Analyse neue Dimensionen, indem sowohl sprachliche wie nicht-sprachliche Aspekte der Kommunikation differenziert erfasst, dokumentiert und somit für Forschungsfragen zugänglich gemacht werden können. Welche Möglichkeit diese quantitativen und qualitativen Zugänge konkret bieten, werden künftige Pilotstudien zeigen.

6. Fraunhofer IAIS Audio Mining-System: Stand und Weiterentwicklung

6.1 Anpassung des Audio Mining Systems für Oral History

Im Idealfall liefert das Audio Mining System eine nahezu perfekte und fehlerfreie Transkription und ersetzt die manuelle Erfassung vollständig. Jedoch ist trotz allen bisherigen Fortschritts in der Spracherkennung bis heute eine derart fehlerfreie automatische Transkription auf menschlichem Niveau für beliebige Sprachaufzeichnungen nicht möglich.

Wie bereits beschrieben, ist das Audio Mining-System mit Rundfunk-Sprachaufzeichnungen trainiert. Diese sind in der Regel professionell unter Einsatz hochwertiger Aufnahmegeräte aufgezeichnet. Die Sprecher artikulieren während der Aufzeichnungen sehr deutlich und klar. Spontansprache und Umgangssprache sind eher eine Ausnahme in den Trainingsdaten. Daher stellen insbesondere Oral History-Interviews das Audio Mining-System vor große Herausforderungen, denn diese weisen oft nur eine vergleichsweise geringe Audio-Qualität auf und sind, bedingt durch die an die Alltagskommunikation angelehnte Form, geprägt von Spontansprache, Umgangssprache und undeutlicher Artikulation.

Ein optimales Spracherkennungsergebnis könnte wahrscheinlich erreicht werden, wenn das Spracherkennungssystem auf einigen tausend Stunden zeitlich alignierter Oral History-Interviews trainiert wird. Dies ist aktuell jedoch auf Grund von mangelnden Trainingsdaten nicht umsetzbar. Um dennoch eine optimierte Spracherkennung auf Oral History-Interviews zu erreichen, werden im Rahmen des KA³-Projekts verschiedene Methoden untersucht, wie die vorhandenen Trainingsdaten aus dem Rundfunkbereich an den Anwendungsfall „Oral History" angepasst werden können.

Im ersten Schritt soll die Robustheit des akustischen Modells gegenüber akustischen Störungen verbessert werden. Die Arten von akustischen Störungen in Oral History-Aufzeichnungen sind zahlreich und sehr unterschiedlich. Die beiden häufigsten akustischen Störungen sind Hintergrundstörgeräusche und Raumhall von kleinen oder mittelgroßen Räumen. Ein solcher Raumhall entsteht durch Reflektionen des Schalls an glatten Wänden des Raumes und Überlagerung am Mikrofon während der Aufzeichnungen. Dies äußert sich besonders stark bei großem Abstand der Sprecher zum Mikrofon, wie beispielsweise beim Einsatz eines Tischmikrofons. Menschliche Zuhörer werden von dieser Art der Störung in der Regel kaum beeinträchtigt. Eine Aufzeichnung mit solchem Raumhall mag sich für einen Zuhörer allenfalls „dumpf" oder „flach" anhören, hat jedoch starken Einfluss auf die spektrale Zusammensetzung des Signals und beeinträchtigt die auf den spektralen Merkmalen basierende Spracherkennung.

Daher werden die Rundfunk-Trainingsdaten mittels sogenannter „Data Augmentation" künstlich verschlechtert, indem aus einem Datensatz zufällig ausgewählte Störgeräusche und Raumhall verschiedener Räume eingefügt werden. Das Ziel besteht darin, dem akustischen Modell beim Training eine möglichst große Menge an unterschiedlich gestörten Sprachaufzeichnungen bereitzustellen, so dass das Modell sich auf die akustischen Merkmale der Sprache verlässt, die robust gegenüber den Störungen sind. Dies wird als Multi-Condition-Training bezeichnet.

6.2 Erkennungsergebnisse unter Einsatz von angepassten Trainingsdaten

Die Qualität der Transkriptionen wird typischerweise in Wortfehlerraten (WER, kurz für „word error rate") angegeben. Die Wortfehlerrate ist das Verhältnis zwischen der minimalen Anzahl an Editieroperationen (einfach gesagt: Korrekturen), die notwendig sind, um die Referenzwortfolge (die korrekte Transkription) in die Hypothesewortfolge (das Ergebnis der Spracherkennung) zu transformieren, und der gesamten Anzahl an Worten der Referenz. Etwas präziser wird die WER definiert als:

WER := (I+S+D) / N, wobei
N die gesamte Anzahl an Worten in der Referenz ist,
I die minimale Anzahl an „Insertions" (zusätzlich einzufügende Worte),
S die minimale Anzahl an „Substitutions" (zu ersetzende Worte),
D die minimale Anzahl an „Deletions" (zu löschende Worte)

Unter Einsatz der im Vorfeld beschriebenen repräsentativen Testmenge an Oral History-Aufzeichnungen, wurden Sprecherkennungsexperimente mit unterschiedlich trainierten akustischen Modellen durchgeführt. Das Standardmodell des Audio Mining-Systems erreicht zu Projektbeginn 2015 auf dieser Testmenge lediglich eine mittlere Wortfehlerrate von 55 %. Unter Einsatz von modernsten Topologien neuronaler Netze in Verbindung mit Multi-Condition-Training konnte bereits in ersten Experimenten

eine Verringerung der mittleren Wortfehlerrate auf unter 40 % erreicht werden, während hierfür, aufgrund des hohen Berechnungsaufwands, zunächst lediglich 128 Stunden anstelle der gesamten 1.000 Stunden Trainingsdaten verwendet wurden. (vgl. Gref/Köhler/Leh 2018). Dies entspricht einer relativen Verbesserung von ca. 27 Prozent. Akustische Modelle, die in weiterführenden Experimente mit 1.000 Stunden Trainingsdaten im Multi-Condition-Setup trainiert wurden, erreichen eine Wortfehlerrate von 29,5 % (vgl. Gref/Schmidt/Köhler 2018).

6.3 Fehlertypen bei der Spracherkennung

Trotz der Verbesserungen bei der Spracherkennung enthalten die Transkripte Erkennungsfehler. Wenngleich die automatische Erkennung durch einen statistischen Prozess und Algorithmus erfolgt und im Einzelnen ein Fehler schwer nachzuvollziehen ist, lassen sich durchaus verschiedene Typen und Ursachen von Erkennungsfehlern kategorisieren. Diese gilt es durch geeignete Maßnahmen zu reduzieren.

Out-of-Vocabulary (OOV): Falls ein zu erkennendes Wort nicht im Lexikon des Spracherkennungssystems vorhanden ist und in dem Interview gesprochen wird, versucht der Erkenner, ein möglichst ähnlich klingendes Wort zu hypothetisieren. In einem Interview erwähnt der Sprecher das Wort „Dünnwald“, ein Stadtteil von Köln. Dieser Ortsname ist jedoch nicht in dem Vokabular des Erkenners, der dann versucht ein phonetisch ähnliches Wort zu erkennen. In diesem Fall entscheidet sich der Erkenner für den Ortsnamen „Grünwald“, ein bekannter Stadtteil von München. Diese Verwechslung führt dann zu einer Fehlerkennung. Das Spracherkennungssystem hat aktuell einen Wortschatz von 1 Million Wörter und deckt damit einen großen Umfang der deutschen Sprache ab. Allerdings sind immer wieder gesprochene Orts- und Eigennamen nicht im Lexikon des Erkenners enthalten. Um die Quote der OOV gering zu halten, muss das Lexikon des Spracherkenners für den jeweiligen Anwendungskontext gegebenenfalls angepasst werden.

Zusammengesetzte Wörter: In der deutschen Sprache können und werden sehr häufig zusammengesetzte Wörter gebildet. Dies kann dazu führen, dass das Spracherkennungssystem versucht, die beiden Wörter in Einzelwörter zu zerlegen (Beispiel: „Problemkonstellation“ in „Problem“ und „Konstellation“). Dies wird bei der Evaluierung nicht nur als ein, sondern sogar als zwei Fehler gewertet (eine Substitution und eine Deletions), wenngleich die Einzelwörter richtig erkannt wurden. Außerdem können die Wörter nahezu beliebig zusammengesetzt werden und wiederum zu den bereits beschriebenen Out-Of-Vocabulary-Effekten führen. In einem Interview wurde beispielsweise das Wort „Kriegerwitwensöhne“ verwendet, das dem Spracherkenner nicht bekannt war. In diesem Fall erzeugt der Spracherkenner die fehlerhafte Ausgabe „Krieger Witwen Söhne“.

Versprecher und undeutliche Aussprache: Das Spracherkennungssystem versucht in der Regel, die Standardaussprache zu erkennen. In den Interviews werden aber oftmals Wörter unvollständig ausgesprochen oder auch abgebrochen. Je nachdem wie die Referenztranskription erstellt wurde, entstehen dadurch bei der Evaluierung Erkennungsfehler.

Darüber hinaus besteht bei der Spracherkennung die Herausforderung, dass sich das System für phonetisch ähnliche Wörter entscheidet. Daher ist es besonders wichtig, für jedes Wort eine möglichst exakte phonetische Transkription vorab erzeugt zu haben.

6.4 Automatische Alignierung

Für viele Oral History-Interviews sind durch etliche Stunden mühsamer Arbeit bereits manuell Transkripte angefertigt worden, die eine höhere Qualität als die aktuelle Spracherkennung aufweisen. Jedoch fehlt oft der zeitliche Zusammenhang zwischen dem Audiosignal und dem Transkript, um beide zielführend, beispielsweise durch Untertitelung, einzusetzen. Forced Alignment ist ein auf der Spracherkennung basierendes Verfahren, das diesen zeitlichen Zusammenhang wiederherstellen kann. Beim Forced Alignment wird eine Art „erzwungene" Erkennung mit dem Spracherkennungssystem durchgeführt, bei dem lediglich das gegebene Referenztranskript als erlaubte Wortfolge zugelassen wird. Das Verfahren bestimmt anschließend den Pfad mit den wahrscheinlichsten Zeitmarken für das Transkript und erlaubt im Idealfall eine wortgenaue zeitliche Zuordnung.

Entscheidend für die Qualität der Alignierung ist unter anderem die Qualität des Transkripts. Fehlerhafte Annotationen und zusätzliche, vor der Alignierung nicht entfernte Einfügungen im Transkript werden vom Algorithmus nicht erkannt und ebenfalls versucht, auf das Audiosignal abzubilden. Mit zunehmender Länge des Signals steigen die Komplexität des Verfahrens und die Fehleranfälligkeit. Während die Alignierung eines Referenztranskripts von wenigen Sekunden Länge in der Regel keine große Herausforderung darstellt, können bei mehrere Stunden dauernden Aufzeichnungen bereits durch einige falsch alignierte Worte signifikante Abweichungen zwischen Text und Audio entstehen. Da Oral History-Interviews in der Regel eine ebensolche Länge aufweisen und oft nicht weiter unterteilt sind, wird für diesen speziellen Anwendungsfall im Rahmen des KA³-Projekts an Modifikationen und Anpassungen des Algorithmus gearbeitet.

Abgesehen von diesem Anwendungsfall des Forced Alignment, könnten die zeitlich alignierten Transkripte als neue Trainingsdaten für das Spracherkennungssystem verwendet werden. Hierbei ist jedoch die Zielsetzung etwas abweichend vom obigen Anwendungsfall. Während für eine Untertitelung versucht wird, das gesamte Transkript auf ein gegebenes Audiosignal abzubilden, und hierbei kleine Abweichungen in Kauf genommen werden, ist für die Erzeugung von Trainingsdaten eine hohe Konfidenz bei der Alignierung entscheidend. Hierbei wird daher erlaubt, dass ein Teil des Transkripts verworfen wird, der nicht sicher aligniert werden kann.

7. Fazit und Ausblick

Das Fraunhofer IAIS Audio Mining birgt ein großes Potential für Oral History-Archive. Durch die automatische Analyse und Transkription von Interviews unter Einsatz verschiedener Sprachanalysealgorithmen ermöglicht das System unter anderem die Recherche in großen Datenbeständen bei direktem Zugriff auf das Audiosignal, die Erschließung von nicht händisch transkribierten Interviews sowie Sekundäranalysen von Interviews mittels weiterführenden Ansätzen, beispielsweise die quantitative Analysen von mehreren hundert Interviews. Für all dies ist insbesondere die Leistungsfähigkeit der Spracherkennung von höchster Bedeutung.

Im Rahmen des KA³-Projekts konnte eine deutliche Verbesserung der Spracherkennungsqualität für die herausfordernden Oral History-Interviews erreicht werden. Aktuell beträgt die durchschnittliche Wortfehlerrate auf einer repräsentativen Testmenge von Aufzeichnungen des Hagener Oral History-Korpus 29,5 %, was einer relativen

Verbesserung von ca. 46 % gegenüber dem ursprünglichen System entspricht. Für neuere Aufnahmen, die mit professioneller Mikrofontechnik aufgezeichnet wurden, liegen die Wortfehlerraten deutlich darunter. Aus anderen Anwendungsdomänen besteht die Erkenntnis, dass Wortfehlerraten von ca. 20 bis 25 % bereits zu sehr guten Suchergebnissen führen und die Audio Mining-Technologie sinnvoll für Recherchezwecke eingesetzt werden kann. Darüber hinaus bietet sich das Verfahren zur automatischen zeitlichen Alignierung an, um mit vorhandenen Transkriptionstexten eine zeitliche Zuordnung des gesprochenen und bereits erstellten Textes zu erreichen. Grundsätzlich kann dieses Forced-Alignment entweder über das Fraunhofer IAIS Audio Mining-System oder über den WebMaus Services des Bavarian Archive for Speech Signals (BAS)[7] genutzt werden.

Im weiteren Verlauf des Forschungsprojektes wird an der kontinuierlichen Verbesserung der Spracherkennung für Oral History-Interviews geforscht und beispielsweise untersucht, wie sich bereits vorhandene, transkribierte Interviews für die weitere Verbesserung der Spracherkennung nutzbar machen lassen. Weiteres Verbesserungspotential besteht in der Verwendung von Texten für die Sprachmodellierung, die zu den Themen und Kontexten der Oral History-Interviews passen. Dies führt in der Regel zu einer deutlichen Reduktion der Out-Of-Vocabluary-Quote. Außerdem werden die Sprachmodelle so angepasst, dass die Themen und Wortfolgen der Interviewten besser abgebildet werden. Mit diesen Verbesserungen wird angestrebt, die Fehlerrate nochmals deutlich zu senken, so dass die von der Spracherkennung automatisch generierten Texte wie ein händisch erstelltes Transkript lesbar sind.

Jenseits der Verbesserung der Spracherkennung für Oral History-Interviews gilt es, weitere Informationen aus dem Sprachsignal zu extrahieren. Sprecherwechsel, Backchannel-Ereignisse („ähm“, „ja“, „hmm“), Sprecherinformationen und Emotionen sind weitere wichtige Metadaten, die ein Oral History-Interview beschreiben. Die Analyse von Interaktionsmustern kann neue Forschungsansätze in der Oral History-Forschung aufzeigen. Aktuell sind dazu erste Untersuchungen gestartet. Allerdings bedarf es hier weiterer Forschungsanstrengungen, robuste Verfahren und Anwendungen zu entwickeln und den Geisteswissenschaften zur Verfügung zu stellen.

Literatur

Apel, Linde (2011): Mündliche Quellen in der Werkstatt der Erinnerung, in: Linde Apel, Stefanie Schüler-Springorum, Klaus David (Hg.): Aus Hamburg in alle Welt. Lebensgeschichten jüdischer Verfolgter in der Werkstatt der Erinnerung, München/Hamburg 2011, 201-218.

Apel, Linde (2015): Oral History reloaded. Zur Zweitauswertung von mündlichen Quellen, in: Westfälische Forschungen. Zeitschrift des LWL-Instituts für westfälische Regionalgeschich-te, hrsg. von Bernd Walter und Thomas Küster, 65, 243-254.

Baum, Doris, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler und Joachim Köhler (2010): DiSCo - A German Evaluation Corpus for Challenging Problems in the Broadcast Domain, in: Nicoletta Calzolari (Conference Chair), et al.(Hg.): Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta.

Cheng, Gaofeng, Vijayaditya Peddinti, Daniel Povey, Vimal Manohar, Sanjeev Khudanpur, und Yonghong Yan (2017): An Exploration of Dropout with LSTMs, in: Proc. Interspeech 2017, 1586-1590. https://doi.org/10.21437/Interspeech.2017-129

7 https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface.

Gref, Michael, Christoph Schmidt und Joachim Köhler (2018): Improving Robust Speech Recognition for German Oral History Interviews Using Multi-Condition Training, in: Speech Communication; 13. ITG Symposium, Oldenburg, Germany, 13, 256-260.

Gref, Michael, Joachim Köhler und Almut Leh (2018): Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research, in: Nicoletta Calzolari (Conference Chair), et al. (Hg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan.

Hochreiter, Sepp, und Jürgen Schmidhuber (1997): Long Short-Term Memory, in: Neural Computation (MIT Press) 9, Nr. 8 (November 1997), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Leh, Almut (2015): Vierzig Jahre Oral History in Deutschland. Betrag zu einer Gegenwartsdiagnose von Zeitzeugenarchiven am Beispiel des Archivs „Deutsches Gedächtnis“, in: Westfälische Forschungen. Zeitschrift des LWL-Instituts für westfälische Regionalgeschichte, hrsg. von Bernd Walter und Thomas Küster, 65, 255-268.

Leh, Almut und Doris Tausendfreund (2011): Archiving Audio and Video Interviews, in: Carlos Nunes Silva (Hg.): Online Research Methods in Urban and Planning Studies: Design and Outcomes, Hershey (PA, USA), 353-367.

Niethammer, Lutz (Hg.) (1983a): „Die Jahre weiß man nicht, wo man die heute hinsetzen soll.“ Faschismuser-fahrungen im Ruhrgebiet, Berlin/Bonn.

Niethammer, Lutz (Hg.) (1983b): „Hinterher merkt man, daß es richtig war, daß es schiefgegangen ist.“ Nachkriegserfahrungen im Ruhrgebiet, Berlin/Bonn.

Niethammer, Lutz und Alexander von Plato (Hg.) (1985): „Wir kriegen jetzt andere Zeiten.“ Auf der Suche nach der Erfahrung des Volkes in nachfaschistischen Ländern, Berlin/Bonn.

Oard, Doug (2012): Can Automatic Speech Recognition Replace Manual Transcription?, in: Doug Boyd, Steve Cohen, Brad Rakers und Dean Rehberger (Hg.): Oral History in the Digital Age. Washington, D.C.: Institute of Museum and Library Services.

Peddinti, Vijayaditya, Daniel Povey, und Sanjeev Khudanpur (2015): A time delay neural network architecture for efficient modeling of long temporal contexts, in: {INTERSPEECH} 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015, 3214-3218.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer und Karel Vesely (2011): The Kaldi Speech Recognition Toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.

Sak, Hasim, Andrew W. Senior und Françoise Beaufays (2014); Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, in: CoRR abs/1402.1128.

Schmidt, Christoph Andreas, Michael Stadtschnitzer und Joachim Köhler (2016): The Fraunhofer IAIS Audio Mining System: Current State and Future Directions, in: Speech Communication; 12. ITG Symposium, Paderborn, Germany, 12, 115-119.

Schütze, Fritz (1976): Zur Hervorlockung und Analyse von Erzählungen thematisch relevanter Geschichten im Rahmen soziologischer Feldforschung – dargestellt an einem Projekt zur Erforschung von kommunalen Machtstrukturen. In: Arbeitsgruppe Bielefelder Soziologen (Hg.): Kommunikative Sozialforschung. München, 159-260.

Stadtschnitzer, Michael, Jochen Schwenninger, Daniel Stein und Joachim Köhler (2014): Exploiting the Large-Scale German Broadcast Corpus to Boost the Fraunhofer IAIS Speech Recognition System, in: Nicoletta Calzolari (Conference Chair), et al. (Hg.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland.

Thomas William, Isaac und Florian Znaniecki (1980/1920): The Polish Peasant in Europe and America, Boston.

Tritschler, Alain und Ramesh Gopinath (1999): Improved speaker segmentation and segments clustering using the bayesian information criterion, in: EUROSPEECH'99. Budapest, Hungary: ISCA.

Waibel, A., T. Hanazawa, G. Hinton, K. Shikano und K. J. Lang (1989): Phoneme recognition using time-delay neural networks, in: IEEE Transactions on Acoustics, Speech, and Signal Processing 37, Nr. 3 (Mar 1989), 328-339. https://doi.org/10.1109/29.21701

Zusammenfassung

Dieser Beitrag beleuchtet die Möglichkeiten und die Herausforderungen der Audio Mining-Technologie für die automatisierte Transkription von Oral History-Interviews. Durch die erheblichen Fortschritte in der Spracherkennung deutet sich ein sinnvoller Einsatz der Technologie in den Geisteswissenschaften zur Transkription von Interviews an. Dies eröffnet eine Reihe von Perspektiven für die interviewbasierte Forschung. Erstens lassen sich aufwendige und kostenintensive Transkriptionsarbeiten reduzieren, zweitens ist die Tonspur auf Wortebene direkt mit dem Transkript verbunden und über eine Suchanwendung zugänglich, drittens können weitaus größere Mengen an Interviews recherchiert und ausgewertet werden.

Auf der anderen Seite stellen Oral History-Interviews, vor allem ältere Aufnahmen, hinsichtlich Aufnahmequalität und spontaner sowie dialektaler Sprechweisen eine erhebliche Herausforderung dar, so dass aktuell Forschungsarbeiten notwendig sind, um die Leistungsfähigkeit der Spracherkennung auf notwendige Qualität zu heben. Diese Forschungs- und Entwicklungsarbeiten sind Gegenstand des vom BMBF geförderten Projektes KA³ (Kölner Zentrum Analyse und Archivierung von AV-Daten). Dieser Beitrag gibt eine Übersicht über die eingesetzten Technologien zur Sprachanalyse, die Funktionsweise des Fraunhofer IAIS Audio Mining-Systems, das Oral History-Archiv „Deutsches Gedächtnis“ der FernUniversität in Hagen, die aktuell erzielten Ergebnisse sowie aktuelle Forschungsansätze zur Verbesserung des Systems.